Pathologist-Level Grading of Prostate Biopsies with Artificial Intelligence

Peter Ström, M.Sc.[1]*, Kimmo Kartasalo, M.Sc.[2]*, Henrik Olsson, M.Sc.[1], Leslie Solorzano, M.Sc.[3], Brett Delahunt, M.D.[4], Daniel M Berney, M.D.[5], David G Bostwick, M.D.[6], Andrew J. Evans, M.D.[7], David J Grignon, M.D.[8], Peter A Humphrey, M.D.[9], Kenneth A Iczkowski, M.D.[10], James G Kench, M.D.[11], Glen Kristiansen, M.D.[12], Theodorus H van der Kwast, M.D.[7], Katia RM Leite, M.D.[13], Jesse K McKenney, M.D.[14], Jon Oxley, M.D.[15], Chin-Chen Pan, M.D.[16], Hemamali Samaratunga, M.D.[17], John R Srigley, M.D.[18], Hiroyuki Takahashi, M.D.[19], Toyonori Tsuzuki, M.D.[20], Murali Varma, M.D.[21], Ming Zhou, M.D.[22], Johan Lindberg, Ph.D[1], Cecilia Bergström, Ph.D [23], Pekka Ruusuvuori, Ph.D [2], Carolina Wählby, Ph.D [3,24], Henrik Grönberg, M.D.[1,25], Mattias Rantalainen, Ph.D [1], Lars Egevad, M.D.[26], and Martin Eklund, Ph.D [1]

** Both authors contributed equally to this study.*

Corresponding author: Dr. Martin Eklund; Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, SE-171 77 Stockholm, Sweden; martin.eklund@ki.se; +46 737121611

1. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
2. Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland.
3. Centre for Image Analysis, Dept. of Information Technology, Uppsala University, Uppsala, Sweden.
4. Department of Pathology and Molecular Medicine, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand.
5. Barts Cancer Institute, Queen Mary University of London, London, UK.
6. Bostwick Laboratories, Orlando, FL, USA.
7. Laboratory Medicine Program, University Health Network, Toronto General Hospital, Toronto, ON, Canada.
8. Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA.
9. Department of Pathology, Yale University School of Medicine, New Haven, CT, USA.
10. Department of Pathology, Medical College of Wisconsin, Milwaukee, WI, USA.
11. Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital and Central Clinical School, University of Sydney, Sydney, NSW, Australia.
12. Institute of Pathology, University Hospital Bonn, Bonn, Germany.
13. Department of Urology, Laboratory of Medical Research, University of São Paulo Medical School, São Paulo, Brazil.
14. Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA.
15. Department of Cellular Pathology, Southmead Hospital, Bristol, UK.
16. Department of Pathology, Taipei Veterans General Hospital, Taipei, Taiwan.
17. Aquesta Uropathology and University of Queensland, Brisbane, Qld, Australia.
18. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada.
19. Department of Pathology, Jikei University School of Medicine, Tokyo, Japan.
20. Department of Surgical Pathology, School of Medicine, Aichi Medical University, Nagoya, Japan.
21. Department of Cellular Pathology, University Hospital of Wales, Cardiff, UK.
22. Department of Pathology, UT Southwestern Medical Center, Dallas, TX, USA.
23. Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden.
24. BioImage Informatics Facility of SciLifeLab, Uppsala, Sweden.
25. Department of Oncology, S:t Göran Hospital, Stockholm, Sweden.
26. Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden.

Abstract

Background: An increasing volume of prostate biopsies and a world-wide shortage of uro-pathologists puts a strain on pathology departments. Additionally, the high intra- and inter-observer variability in grading can result in over- and undertreatment of prostate cancer. Artificial intelligence (AI) methods may alleviate these problems by assisting the pathologist to reduce workload and harmonize grading.

Methods: We digitized 6,682 needle biopsies from 976 participants in the population based STHLM3 diagnostic study to train deep neural networks for assessing prostate biopsies. The networks were evaluated by predicting the presence, extent, and Gleason grade of malignant tissue for an independent test set comprising 1,631 biopsies from 245 men as well as an external validation set of 330 biopsies from 73 men. We additionally evaluated grading performance on 87 biopsies individually graded by 23 experienced urological pathologists from the International Society of Urological Pathology. We assessed discriminatory performance by receiver operating characteristics (ROC) and tumor extent predictions by correlating predicted millimeter cancer length against measurements by the reporting pathologist. We quantified the concordance between grades assigned by the AI and the expert urological pathologists using Cohen's kappa.

Results: The AI achieved an area under the ROC curve of 0·997 for distinguishing between benign and malignant biopsy cores on the independent test set and 0·986 on the external validation set. The correlation between millimeter cancer predicted by the AI and assigned by the reporting pathologist was 0·96 for the independent test set and 0·87 for the external validation set. For assigning Gleason grades, the AI achieved an average pairwise kappa of 0·62. This was within the range of the corresponding values for the expert pathologists (0·60 to 0·73).

Conclusions: The performance of the AI to detect and grade cancer in prostate needle biopsy samples was comparable to that of international experts in prostate pathology. AI has potential to reduce high intra-observer variability and to provide diagnostic expertise in regions where this is currently unavailable.

Introduction

Histopathological evaluation of prostate biopsies is critical to the clinical management of men suspected of having prostate cancer. Despite this importance, the histopathological diagnosis of prostate cancer is associated with several challenges:

- More than one million men undergo prostate biopsy in the United States annually.[1] With the standard biopsy procedure resulting in 10-12 needle cores per patient, more than 10 million tissue samples need to be examined by pathologists. The increasing incidence of prostate cancer in an aging population means that the number of biopsies is likely to further increase.
- It is recognized that there is a shortage of pathologists internationally. In China, there is only one pathologist per 130,000 population, while in many African countries the ratio is of the order of one per million.[2,3] Western countries are facing similar problems, with an expected decline in the number of practicing pathologists due to retirement.[4]
- Gleason grade is the most important prognostic factor for prostate cancer and is crucial for treatment decisions. Gleason grade is based on morphologic examination and is recognized to be notoriously subjective. This is reflected in high intra- and inter-pathologist variability in reported grades, as well as both under- and over-diagnosis of prostate cancer.[5,6]

A possible solution to these challenges is the application of artificial intelligence (AI) to prostate cancer histopathology. The development of an AI to identify benign biopsies with high accuracy would decrease the workload of pathologists and allow them to focus on difficult cases. Further, an accurate AI could assist the pathologist with the identification, localization and grading of prostate cancer among those biopsies not culled in the initial screening process, thus providing a safety net to protect against potential misclassification of biopsies. AI-assisted pathology assessment could harmonize grading and reduce–inter-observer variability, leading to more consistent and reliable diagnoses and better treatment decisions.

Using high resolution scanning, tissue samples can be digitized to whole slide images (WSI) and utilized as input for the training of deep neural networks (DNN), an AI technique which has been successful in many fields, including medical imaging.[7–10] Despite the many successes of AI, little work has been undertaken in prostate diagnostic histopathology.[11–16] Attempts at grading prostate biopsies by DNNs have been limited to small datasets or subsets

of Gleason patterns, and they have lacked analyses of the clinical implications of the introduction of AI-assisted prostate pathology.

In this study, we aimed to develop an AI with clinically acceptable accuracy for prostate cancer detection, localization, and Gleason grading. To achieve this, we digitized 8,313 samples from 1,222 men included in the prospective and population based STHLM3 prostate cancer diagnostic study undertaken in 2012-2015.[17,18] We evaluated the performance of the model on an independent test set as well as an external validation set (external lab and scanner), and through a comparison with 87 cases of prostate cancer graded by the International Society of Urological Pathology (ISUP) Imagebase panel consisting of 23 experienced uro-pathologists.[19]

Methods

Sample population and data collection

Between 2012 and 2015, the prospective and population-based STHLM3 study evaluated a diagnostic model for prostate cancer in men aged between 50 and 69 years.[17,18] Among the 59,159 participants, 7,406 (12·5%) underwent systematic biopsy according to a standardized protocol consisting of 10 or 12 needle cores; with 12 cores being taken from prostates larger than 35 cm^3 (Table 1). Urologists who participated in the study and the study pathologist were blinded to the clinical characteristics of the patients. A single pathologist (L.E.) graded all biopsy cores according to the ISUP grading classification (where Gleason scores 6, 3+4=7, 4+3=7, 8, and 9-10 are reported as ISUP grade 1 to 5, also referred to as Gleason Grade Groups). L.E. also delineated cancerous areas using a marker pen and measured the linear cancer extent.[20,21]

The biopsy cores were formalin fixed and stained with hematoxylin and eosin. A selection of 8,313 biopsies from 1,222 STHLM3 participants was digitized. The cases were chosen to represent the full range of diagnoses, with an over-representation of high-grade disease. To further enrich the data with high-grade cases, 271 slides from 93 men with ISUP 4 and 5 prostate cancers were obtained from outside STHLM3 (see Appendix for details). These slides were re-graded by L.E., digitized and utilized for training purposes only. We used 1,631 cores from a random selection of 246 (20%) men to evaluate the performance of the AI (the "independent test set"), while the rest were used for model training. That is, *all* biopsies from a given man were assigned to either the training or the test dataset.[22]

Since slides from different pathology labs differ in appearance and quality due to differences in slide preparation and since WSI characteristics and appearance vary by scanner, it is crucial to assess the performance of DNN models on external labs and scanners (i.e. images of slides from different pathology labs and scanners than the images on which the model was trained) from a real-world clinical setting. We therefore obtained 330 slides (73 men) from the Karolinska University Hospital and digitized them on the scanner available at the Karolinska University Hospital pathology lab to replicate their entire workflow of lab processing and slide digitization (the "external validation set"). The selection of slides was enriched for higher ISUP grades to permit evaluation of predictions for these uncommon grades (Table 1). L.E. graded all biopsies in the external test set to avoid confoundment between introducing a different reporting pathologist and a different lab and scanner workflow simultaneously.

As an additional test set, we digitized 87 cores from the Pathology Imagebase, a reference database launched by ISUP to promote the standardization of reporting of urological pathology.[19] These cases were independently reviewed by 23 highly experienced urological pathologists (The ISUP Imagebase panel). Cores from the men in the three test sets were not part of model development and were excluded from any analysis until the final evaluation.

The study protocol was approved by Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3 and 2018/845–32). For details concerning data collection, see Appendix.

Artificial intelligence framework

Image pre-processing

We processed the WSIs with a segmentation algorithm based on Laplacian filtering to identify the regions corresponding to tissue sections and annotations drawn adjacent to the tissue (Figure S1). We then extracted digital pixel-wise annotations, indicating the locations of cancerous tissue of any grade, by identifying the tissue region corresponding to each annotation. To obtain training data representing the morphological characteristics of Gleason patterns 3, 4 and 5, we extracted numerous partially overlapping smaller images, or *patches*, from each WSI. Each patch was small enough to largely represent only benign or cancerous tissue. We used patch dimensions of 598 x 598 pixels (approx. 540 x 540 µm) at a resolution corresponding to 10X magnification (pixel size approx. 0·90 µm). The process resulted in approximately 5·1 million patches usable for training a DNN. See Appendix for details (Table S1).

Deep neural network model for classification of image patches

We used two convolutional DNN ensembles, each consisting of 30 Inception V3 models pre-trained on ImageNet, with classification layers adapted to our outcome.[23,24] The first ensemble performed binary classification of image patches into benign or malignant, while the second ensemble classified patches into Gleason patterns 3 to 5. To reduce label noise in the latter case, we trained the ensemble on patches extracted from cores containing only one Gleason pattern (i.e. cores with Gleason score 3+3, 4+4, or 5+5). Importantly, the test data still contained cores of all grades to provide a real-world scenario for evaluation. Each DNN in the first and the second ensemble thus predicted the probability of each patch being malignant, and whether it represented Gleason pattern 3, 4, or 5, respectively. See Appendix for details (Figure S2).

Boosted tree model for core-level estimation of cancer grade and length

Once the probabilities for the Gleason pattern at each location of the biopsy core were obtained from the DNN ensembles, we mapped them to core-specific characteristics (ISUP grade and cancer length) using boosted trees.[25] All cores in the training data were used for training the boosted trees. Specifically, aggregated features from the patch-wise probabilities predicted by each DNN for each core were used as input to the boosted trees, and the clinical assessment of ISUP score and cancer length were used as outcomes. The ISUP grade group was assigned based on a Bayesian decision rule of the core-level classifier to obtain ISUP predictions at a clinically relevant operating point (see Appendix for details).

Model interpretation

With the aim of interpreting the representation of the image data learned by the DNN models, we performed a visualization of the feature space. To obtain the feature representation of a patch, we extracted the activations of the DNN's penultimate layer for the patch in question. To allow visualization of the high-dimensional feature space, we performed dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE)[26]. Additionally, in order to gain insights on the features of the input space the DNN model bases its decisions on, we applied the deep Taylor decomposition approach implemented in the iNNvestigate toolbox.[27,28] This technique relies on modelling the decision process of a DNN by backtracking the signals observed in response to a particular input image, resulting in a pixel-wise estimate indicating which parts of the input image are most likely to contribute to the DNN's decision-making. See Appendix for details (Figure S8).

Evaluation metrics

Cancer detection

We summarized the operating characteristics of the AI system in a Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC). We then specified a range of acceptable sensitivities for potential clinical use, and evaluated achieved specificity (both on core-level and patient-level) when compared to the pathology report. The enrichment of high-grade disease in the independent test data and the external validation data could potentially inflate the estimated AUC values, since these grades may be easier to discriminate from benign cases compared to e.g. ISUP 1 and 2. Therefore, we also estimated the AUC when ISUP 3-5 cases were removed from the independent test set and the external validation.

Cancer length estimation

We predicted cancer length in each core and compared it to the cancer length described in the pathology report. The comparison was undertaken on individual cores as well as on aggregated cores (i.e. total cancer length) for each man. Linear correlation was assessed on both all cores and men, as well as restricted to positive cores and men.

ISUP grading

Cohen's kappa with linear weights was used for evaluating the AI's performance against the 23 experienced uro-pathologists on the Imagebase test set. Linear weights emphasize a higher level of disagreement of ratings further away from each other on the ordinal ISUP scale, in accordance with previous publications on the Imagebase study.[19] Each of the 87 slides in Imagebase was graded by each of the 23 Imagebase panel pathologists, and additionally by the AI. To evaluate how well the AI agreed with the pathologists, we calculated all pair-wise kappas and summarized the average for each of the 23 raters. In addition, we estimated the kappa with a grouping of the Gleason scores in ISUP grades (grade groups) 1, 2-3 and 4-5. We also estimated Cohen's kappa against the study pathologist's ISUP grading on the independent test set and the external validation set. For the external validation set, we also estimated Cohen's kappa after calibrating the probabilities (i.e. scaling the ISUP probabilities before assigning the predicted class). This was done to investigate whether simple fine tuning (without retraining of the DNN) is likely to suffice to counteract any drop in performance on external data.

Role of the funding source

The funders had no role in study design, data collection, analysis and interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Cancer detection

We estimated the AUC representing the ability of the AI to distinguish malignant from benign cores to 0·997 for the independent test set and 0·986 for the external validation set (Figure 1). The AUC values changed only marginally when ISUP 3-5 cases were removed: from 0.997 to 0.996 for the independent test set and from 0.986 to 0.980 for the external validation data. As an example, at a sensitivity of 99·6% on the independent test set, the AI achieved a specificity of 86·6% (Figure 1; top panel, second row from the top). At this sensitivity level, the AI failed to detect three cores with cancer (two ISUP grade 1 and one ISUP grade 2, all with less than 0·5 mm cancer) across 721 malignant biopsy cores in the independent test data. No cancer was misdiagnosed since other malignant cores from the same men were correctly classified. For predicting whether a man had cancer or not, the AUC was 0·999 and 0·979, respectively.

Cancer length estimation

A visualization of the estimated localization of malignant tissue for an example biopsy is presented in Figure 2B. The correlation between the cancer length estimates of the AI and the measurements of the pathologist was 0·96 (0·93 for positive cores) for the independent test set and 0·87 (0·80 for positive cores) for the external validation set. When aggregating the cancer extent of all cores within a case, the correlation was 0·98 on the independent test set and 0·94 on the external validation set, both for all men and for men positive for cancer (Figure 3). An online tool (https://tissuumaps.research.it.uu.se/sthlm3/) allows for interactive examination of predictions generated for 30 cores randomly selected (5 per ISUP score and 5 benign) from the independent test set and from the external validation set, respectively.

ISUP grading

The average pairwise kappa achieved by the AI on the 87 Imagebase cases was 0·62. The pathologists had values ranging from 0·60 to 0·73, with the study pathologist (L.E.) having a kappa of 0·73. When considering a narrower grouping of ISUP grades (ISUP 1, ISUP 2-3 and

ISUP 4-5), which often forms the basis for primary treatment selection, the AI scored even higher relative to the pathologists (Figure 4A). The grades assigned by the panel and the AI to each Imagebase case are shown in Appendix (Figure S3).

The kappa obtained by the AI relative to the pathology report in the independent test set of 1,631 cores was 0·83 for all cores and 0·70 for positive cores only (Figure 4B). The kappa was 0·70 for all cores and 0·61 for positive cores on the external validation set (Figure 4C). By scaling the ISUP probabilities before assigning the predicted class (calibrating to the new site), the kappa increased to 0·76 for all cores and 0·66 for positive cores on the external validation data (Figure 4D).

Model interpretation

A visualization of the feature space learned by the model to represent the histological image data is shown in Figure 2A. Examination of this visualization reveals that the model has learnt a representation that appears logical and unsurprising – most patches representing benign tissue cluster together, and a smooth transition with increasing malignancy can be observed first from the benign patches towards patches representing Gleason pattern 3, and further towards the cluster of patches labeled as Gleason pattern 4. Interestingly, patches representing Gleason pattern 5 appear as a separate cluster in the feature space. Furthermore, the presence of patch-level label noise is observable, as a number of patches labeled as malignant appear in the cluster of predominantly benign patches, and a cluster of patches labeled as benign can be seen in a region overlapping with Gleason patterns 3 and 4.

We further identified the pixel-level patterns forming the basis of the DNN's grading decisions (Figure S8). Based on this analysis, the model appears to focus mainly on small glandular structures, especially the luminal parts, as well as on cell nuclei. While these observations may appear obvious to a pathologist, they serve as further confirmation that the DNN has identified features which are genuinely relevant for the diagnostic task, as opposed to for example utilizing artifactual patterns in the data resulting from variation in sample or image processing.[29]

Discussion

Grading prostate cancer can be a difficult procedure due to the complex nature of the score and its derivation. This has also been true for computer algorithms aiming at automating

grading. The challenge is not only to develop an AI for this task, but also to demonstrate that it is consistent with current state-of-the-art diagnosis of prostate histopathology. Here, we have for the first time demonstrated AI-based grading of prostate biopsies on the level of leading urological pathologists represented by the ISUP Imagebase panel.

Due to the poor discriminative ability of the prostate specific antigen test and the systematic biopsy protocol of 10-12 needle cores, which is still in common usage, most biopsies encountered in clinical practice are of benign tissue. To reduce the workload of assessing these samples, we evaluated the AI's ability to assist the pathologist by pre-screening benign from malignant cores. With an estimated AUC of 0·997 on the independent test set, the system could automatically remove 809 benign biopsies from 246 men without missing a single man out of the 211 with cancer diagnosed by the study pathologist (Figure 1). Since the pathology report was used as gold standard for this evaluation, the AI, by design, cannot achieve a higher sensitivity than the reporting pathologist. However, the sensitivity of the AI system could in fact be higher, as some malignant cores may be overlooked by the pathologist but detected by the AI. As an illustration of this, Ozkan *et al.* evaluated the agreement of two pathologists in the assessment of cancer in biopsy cores.[5] Following examination of 407 cases, one pathologist found cancer in 231 cases, while the other found cancer in 202 cases. This suggests that an AI can not only streamline the workflow but could also improve sensitivity by detecting cancer foci that would otherwise be accidentally overlooked.

In this study, we have also evaluated the assessment of tumor burden (cancer length). We believe that both cancer detection and cancer length measurements can now be automated without sacrificing patient safety. In support of this and to provide interpretations of the DNN's predictions, we have published on our website high-resolution images of cores randomly selected from the independent test data and the external validation data, accompanied by their ISUP grades and the AI's predictions.

The first attempt to use DNNs for the detection of cancer on prostate biopsies was reported by Litjens *et al.*[15] Using an approach similar to ours but based on a small dataset, they could safely exclude 32% of benign cores. A more recent study by Campanella *et al.* demonstrated an AUC of 0·991 for cancer detection on an independent test set and 0.943 on external validation data.[16] There have also been attempts to undertake grading of prostate tissue derived from prostatectomy or based on tissue microarrays.[14,30] None of these studies achieved expert uro-pathologist level consistency in Gleason grading, estimated tumor burden, or investigated grading on needle biopsies, which is of significance since this is the sampling utilized for diagnosis and grading in virtually every pathology laboratory worldwide.

Moreover, no previous study has used a well-defined cohort of samples to estimate the clinical implications, with respect to key medical operating characteristic metrics such as sensitivity and specificity.[31]

The strengths of our study include the use of well-controlled, prospectively collected and population-based data covering a large random sample of men with both the urologists and the pathologist blinded to patient characteristics. Prostate cancers diagnosed in STHLM3 are representative for a screening-by-invitation setting, and the data include cancer variants that are notoriously difficult to diagnose (pseudohyperplastic and atrophic carcinoma), slides which required immunohistochemistry, mimickers of cancer, slides with thick cuts and fragmented cores and poor staining (Table S6). Despite these difficult cases, the AI achieved near perfect diagnostic concordance with the study pathologist. The study was subjected to a strict protocol, where the splitting of cases into training and test sets was performed at a patient level and all analyses were pre-specified prior to the evaluation of the independent test set, including code for producing tables, figures, and result statistics. A further strength is the use of Imagebase which is a unique dataset for testing the performance of the AI against highly experienced urological pathologists.

We trained the AI using annotations from a single, highly experienced urological pathologist (L.E.). The decision to rely on a single pathologist for model training was done to avoid presenting the AI with conflicting labels for the same morphological patterns and to thereby achieve more consistent predictions. L.E. has in several studies demonstrated high concordance with other experienced uro-pathologists, and therefore represents a good reference for model training.[32,33] For model evaluation, however, it is critical to assess performance against multiple pathologists (Figure 4A).

Several sources of variability affect the AI's predictions. In addition to morphological variability, technical variability is introduced during slide preparation and scanning. Given the sensitivity of DNNs to differences in input data, it is plausible that differences across labs and scanners can invalidate any discriminatory capacity of a DNN.[34] Here, we showed that the capacity of the AI in discriminating between benign and malignant biopsies decreased only marginally on the external validation data compared to the independent test set. We did however observe some reduction in performance with respect to overall Gleason grading (Cohen's kappa decreased from 0.83 to 0.70 from the independent test set to the external validation set with respect to grades assigned by L.E.). This reduction in performance was most notable for ISUP 2 grades (Figure 4C). However, by scaling the AI's predictions for the different classes (i.e. calibrating five scalar parameters to the new site), the results were markedly closer to the

results achieved on the independent test data (Figure 4D). This is a key observation, as it suggests that although some fine tuning to a new site or scanner is likely required to achieve optimal performance, this tuning is lightweight and can be done using little data. Importantly, it does not require redevelopment or retraining of either the DNN models or the slide-level models, which would be infeasible both from a practical and regulatory perspective. Albeit being a limitation of the method, requirement for such calibration is not uncommon when deploying a diagnostic test at a new site (e.g. calibrants are routinely used in laboratory diagnostics to diagnose and prevent site specific differences and drift over time) and is unlikely to present a major hurdle for the clinical application of AI-based diagnostics.

A limitation of this study is the lack of exact pixel-wise annotations, since the annotations may highlight regions that include a mixture of benign and malignant glands of different grades. To address this issue, we trained the algorithm on slides with pure Gleason grades, used a patch size large enough to cover glandular structures but small enough to minimize the presence of mixed grades within a patch, and we focused our attention on core and patient level performance metrics, which avoids caveats of patch-level evaluation and is clinically more meaningful. Another limitation is the difficulty of using a subjective measure like ISUP grade as ground truth for AI models. We approached this problem by evaluating the ISUP grade assigned by the AI against a panel of experienced pathologists. We also confirmed that the classifications of the AI did not substantially differ from the pathologist's when evaluating PSA relapses among the operated men in the trial (see Appendix for details, Table S7). As a consequence of the study design with an enrichment of high-grade disease, the interpretation of an estimated positive predictive value (PPV) is not straightforward (since the PPV is a function of the prevalence in the dataset). We have therefore chosen to not report the PPV, something that we will address in future studies with a design more suitable for estimating the PPV.

Conclusions

We have demonstrated that an AI based on DNNs can grade prostate biopsies at the level of highly experienced urological pathologists and that a DNN's predictions can generalize to external data. We believe that the use of a system like this can increase sensitivity and promote patient safety by providing decision-support and by focusing the attention of the pathologist on regions of interest. In addition, the use of an accurate AI system can reduce pathology workload and high intra-observer variability in the reporting of prostate

histopathology by producing reproducible and consistent grading. A further benefit is that AI can provide diagnostic expertise in regions where this is currently unavailable.

Author contributions

ME had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. PS and KK contributed equally to algorithmic design and implementation, and drafting the manuscript. In addition, PS was mainly responsible for statistical analysis of results and KK was mainly responsible for high-performance computing. HO was mainly responsible for data management and participated in algorithmic design and implementation, and in drafting the manuscript. LS developed the online viewer application allowing visual examination of results. BD was involved in drafting the manuscript. BD, DMB, DGB, LE, AJE, DJG, PAH, KAI, JGK, GK, THVDK, KRML, JKMK, JO, CCP, HS, JRS, HT, TT, MV, MZ performed grading of the ImageBase dataset and provided pathology expertise and feedback. CL was involved in data collection. JL was involved in study design. PR and CW contributed to design and supervision of the study and to algorithmic design. In addition, PR contributed to high-performance computing and CW contributed to designing the online viewer. HG contributed to the conception, design and supervision of the study. MR contributed to the conception, design and supervision of the study and to algorithmic design. LE graded and annotated all the data used in the study, contributed to the conception, design, and supervision of the study, and helped draft the manuscript. ME was responsible for the conception, design and supervision of the study, and contributed to algorithmic design, analysis of results and drafting the manuscript. All authors participated in the critical revision and approval of the manuscript.

Acknowledgements

The Tampere Center for Scientific Computing and CSC - IT Center for Science, Finland are acknowledged for providing computational resources. The S:t Göran Hospital, Stockholm, is acknowledged for providing additional high-grade slides as training data. Carin Cavalli-Björkman, Britt-Marie Hune, Astrid Björklund, and Olof Cavalli-Björkman have been instrumental in logistical handling of the glass slides. Hannu Hakkola, Tomi Häkkinen, Leena Latonen, Kaisa Liimatainen, Teemu Tolonen, Masi Valkonen and Mira Valkonen are acknowledged for their helpful advice. We thank the participants in the Stockholm-3 study for their participation.

Funding

Funding was provided by the Swedish Research Council, Swedish Cancer Society, Swedish Research Council for Health, Working Life, and Welfare (FORTE), Academy of Finland [313921], Cancer Society of Finland, Emil Aaltonen Foundation, Finnish Foundation for Technology Promotion, Industrial Research Fund of Tampere University of Technology, KAUTE Foundation, Orion Research Foundation, Svenska Tekniska Vetenskapsakademien i Finland, Tampere University Foundation, Tampere University graduate school, The Finnish Society of Information Technology and Electronics, TUT on World Tour programme and the European Research Council (grant ERC-2015-CoG 682810). The funders had no role in study design, data collection, analysis and interpretation, writing of the report or making the decision to submit for publication.

Competing interests

HG has five prostate cancer diagnostic related patents pending, has patent applications licensed to Thermo Fisher Scientific, and might receive royalties from sales related to these patents. ME is named on four of these five patent applications. Karolinska Institutet collaborates with Thermo Fisher Scientific in developing the technology for STHLM3. PE, KK, and ME are named on a pending patent related to cancer diagnostics quality control. All other authors declare no competing interests.

References

1 Loeb S, Carter HB, Berndt SI, Ricker W, Schaeffer EM. Complications after prostate biopsy: Data from SEER-Medicare. *J Urol* 2011; **186**: 1830–4.

2 Egevad L, Delahunt B, Samaratunga H, *et al.* The International Society of Urological Pathology Education web-a web-based system for training and testing of pathologists. *Virchows Arch* 2019; published online Feb. DOI:10.1007/s00428-019-02540-w.

3 Adesina A, Chumba D, Nelson AM, *et al.* Improvement of pathology in sub-Saharan Africa. Lancet Oncol. 2013. DOI:10.1016/S1470-2045(12)70598-3.

4 Robboy SJ, Weintraub S, Horvath AE, *et al.* Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med* 2013; **137**: 1723–32.

5 Ozkan TA, Eruyar AT, Cebeci OO, Memik O, Ozcan L, Kuskonmaz I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand J Urol* 2016; **50**:

420–4.

6 Melia J, Moseley R, Ball RY, *et al.* A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology* 2006; **48**: 644–54.

7 Bejnordi BE, Veta M, Van Diest PJ, *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - J Am Med Assoc* 2017; **318**: 2199–210.

8 Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**: 115–8.

9 Silver D, Huang A, Maddison CJ, *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* 2016; **529**: 484–9.

10 Gulshan V, Peng L, Coram M, *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J Am Med Assoc* 2016; **316**: 2402–10.

11 Gummeson A, Arvidsson I, Ohlsson M, *et al.* Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks. In: Medical Imaging 2017: Digital Pathology. 2017: 101400S.

12 Kallen H, Molin J, Heyden A, Lundstrom C, Astrom K. Towards grading gleason score using generically trained deep convolutional neural networks. Proc. - Int. Symp. Biomed. Imaging. 2016; **2016–June**: 1163–7.

13 Jiménez del Toro O, Atzori M, Otálora S, *et al.* Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. Med. Imaging 2017 Digit. Pathol. 2017; **10140**: 101400O.

14 Arvaniti E, Fricker KS, Moret M, *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; **8**: 12054.

15 Litjens G, Sánchez CI, Timofeeva N, *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 2016; **6**: 26286.

16 Campanella G, Hanna MG, Geneslaw L, *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; published online July. DOI:10.1038/s41591-019-0508-1.

17 Grönberg H, Adolfsson J, Aly M, *et al.* Prostate cancer screening in men aged 50-69 years (STHLM3): A prospective population-based diagnostic study. *Lancet Oncol* 2015; **16**: 1667–76.

18 Ström P, Nordström T, Aly M, Egevad L, Grönberg H, Eklund M. The Stockholm-3 Model for Prostate Cancer Detection: Algorithm Update, Biomarker Contribution, and Reflex Test Potential. *Eur Urol* 2018; **74**: 204–10.

19 Egevad L, Delahunt B, Berney DM, *et al.* Utility of Pathology Imagebase for

standardisation of prostate cancer grading. *Histopathology* 2018; **73**: 8–18.

20 Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol* 2016; **40**: 244–52.

21 Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic Gleason grade grouping: Data based on the modified Gleason scoring system. *BJU Int* 2013; **111**: 753–60.

22 Nir G, Karimi D, Goldenberg SL, *et al.* Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images. *JAMA Netw open* 2019; **2**: e190442.

23 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016; **2016–Decem**: 2818–26.

24 Jia Deng, Wei Dong, Socher R, Li-Jia Li, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conf. Comput. Vis. Pattern Recognit. 2009; : 248–55.

25 Chen T, Guestrin C. XGBoost. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16. 2016; : 785–94.

26 van der Maaten L, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *J Mach Learn Res* 2008; **9**: 2579–605.

27 Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit* 2017. DOI:10.1016/j.patcog.2016.11.008.

28 Alber M, Lapuschkin S, Seegerer P, *et al.* iNNvestigate neural networks! *J Mach Learn Res* 2019; **20**: 1–8.

29 Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* 2019; **10**: 1096.

30 Nagpal K, Foote D, Liu Y, *et al.* Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Digit Med* 2019. DOI:10.1038/s41746-019-0112-2.

31 AI diagnostics need attention. *Nature* 2018; **555**: 285.

32 Kweldam CF, Nieboer D, Algaba F, *et al.* Gleason grade 4 prostate adenocarcinoma patterns: an interobserver agreement study among genitourinary pathologists. *Histopathology* 2016; **69**: 441–9.

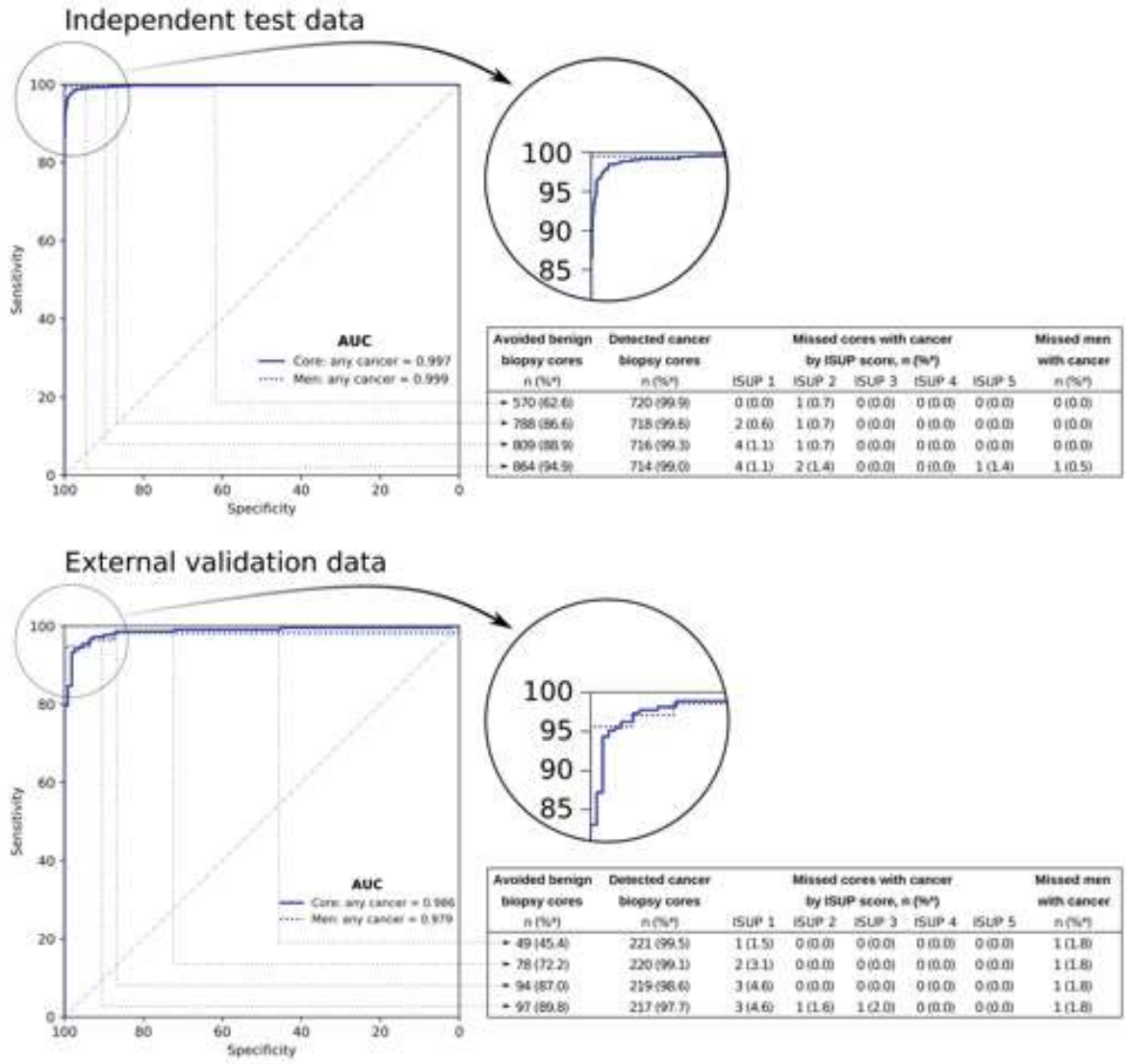
33 Egevad L, Cheville J, Evans AJ, *et al.* Pathology Imagebase—a reference image database for standardization of pathology. *Histopathology* 2017; **71**: 677–85.

34 Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. *CoRR* 2014; **abs/1412.6.**

	Stockholm-3	Digitized (n=1,454)				
	Biopsied (n=7,406)	Training (n=978)	Extra Training (n=93)	Test (n=246)	Imagebase (n=86)	External (n=73)
Per Man	No. (%)	No. (%)	No. (%)	No. (%)	No. (%)	No. (%)
Age†						
<49 yr	45 (0.61)	4 (0.41)	0 (0.0)	1 (0.41)	0 (0.0)	2 (2.74)
50-54 yr	639 (8.63)	76 (7.81)	2 (2.2)	11 (4.47)	10 (11.63)	5 (6.85)
55-59 yr	1221 (16.49)	136 (13.98)	4 (4.3)	44 (17.89)	8 (9.3)	10 (13.7)
60-64 yr	2027 (27.37)	255 (26.21)	8 (8.6)	67 (27.24)	23 (26.74)	12 (16.44)
65-69 yr	3294 (44.48)	482 (49.54)	17 (18.3)	115 (46.75)	44 (51.16)	15 (20.55)
≥70 yr	180 (2.43)	20 (2.06)	57 (61.3)	8 (3.25)	1 (1.16)	29 (39.73)
Missing	0 (0.0)	0 (0.0)	5 (5.4)	0 (0.0)	0 (0.0)	0 (0.0)
Previous negative biopsy						
Yes	505 (6.82)	33 (3.39)	0 (0.0)	13 (5.28)	7 (8.14)	
No	6901 (93.18)	940 (96.61)	0 (0.0)	233 (94.72)	79 (91.86)	
Missing	0 (0.0)	0 (0.0)	93 (100.0)	0 (0.0)	0 (0.0)	
PSA						
<3 ng/mL	1933 (26.1)	228 (23.43)	2 (2.2)	43 (17.48)	13 (15.12)	
3-5 ng/mL	3458 (46.69)	428 (43.99)	2 (2.2)	100 (40.65)	48 (55.81)	
5-10 ng/mL	1612 (21.77)	213 (21.89)	19 (20.4)	73 (29.67)	16 (18.6)	
≥10 ng/mL	403 (5.44)	104 (10.69)	57 (61.3)	30 (12.2)	9 (10.47)	
Missing	0 (0.0)	0 (0.0)	13 (14.0)	0 (0.0)	0 (0.0)	
Digital rectal examination						
Abnormal	680 (9.18)	133 (13.67)	59 (72.0)	39 (15.85)	12 (13.95)	
Normal	6726 (90.82)	840 (86.33)	10 (12.2)	207 (84.15)	74 (86.05)	
Missing	0 (0.0)	0 (0.0)	13 (15.9)	0 (0.0)	0 (0.0)	
Prostate volume‡						
<35 mL	2701 (36.47)	425 (43.68)	26 (28.0)	92 (37.4)	42 (48.84)	
35-50 mL	2494 (33.68)	319 (32.79)	18 (19.4)	82 (33.33)	36 (41.86)	
≥50 mL	2211 (29.85)	229 (23.54)	22 (23.7)	72 (29.27)	8 (9.3)	
Missing	0 (0.0)	0 (0.0)	27 (29.0)	0 (0.0)	0 (0.0)	
Cancer length						
No cancer	4605 (62.18)	139 (14.29)	0 (0.0)	35 (14.23)	0 (0.0)	16 (21.92)
0-1 mm	545 (7.36)	133 (13.67)	2 (2.2)	35 (14.23)	4 (4.65)	1 (1.37)
1-5 mm	922 (12.45)	258 (26.52)	10 (10.8)	61 (24.8)	20 (23.26)	10 (13.7)
5-10 mm	449 (6.06)	135 (13.87)	17 (18.3)	28 (11.38)	20 (23.26)	6 (8.22)
>10 mm	885 (11.95)	308 (31.65)	64 (68.8)	87 (35.37)	42 (48.84)	40 (54.79)
Cancer grade						
Benign	4605 (62.18)	139 (14.29)	0 (0.0)	35 (14.23)		16 (21.92)
ISUP 1 (3 + 3)	1558 (21.04)	413 (42.45)	1 (1.1)	104 (42.28)		12 (16.44)
ISUP 2 (3 + 4)	761 (10.28)	200 (20.55)	1 (1.1)	53 (21.54)		12 (16.44)
ISUP 3 (4 + 3)	253 (3.42)	96 (9.87)	1 (1.1)	16 (6.5)		16 (21.92)
ISUP 4 (4 +4, 3 +5 and 5 + 3)	101 (1.36)	63 (6.47)	19 (20.4)	21 (8.54)		8 (10.96)
ISUP 5 (4 +5, 5 +4 and 5 +5)	128 (1.73)	62 (6.37)	71 (76.3)	17 (6.91)		9 (12.33)

Per Biopsy core	Biopsied (n=83,470)	Training (n=6,662)	Extra Training (n=271)	Test (n=1,821)	Imagebase (n=87)	External (n=330)
Cancer length						
No cancer	73595 (88.17)	3724 (55.73)	1 (0.37)	910 (55.79)	0 (0.0)	108 (32.73)
0-1 mm	3307 (3.96)	915 (13.69)	7 (2.58)	203 (12.45)	8 (9.2)	33 (10.0)
1-5 mm	4135 (4.95)	1239 (18.54)	41 (15.13)	295 (18.09)	44 (50.57)	77 (23.33)
5-10 mm	1822 (2.18)	591 (8.84)	85 (31.37)	150 (9.2)	24 (27.59)	75 (22.73)
>10 mm	611 (0.73)	213 (3.19)	111 (40.96)	73 (4.48)	11 (12.64)	37 (11.21)
Missing	0 (0.0)	0 (0.0)	26 (9.59)	0 (0.0)	0 (0.0)	0 (0.0)
Cancer grade						
Benign	73595 (88.17)	3724 (55.73)	1 (0.37)	910 (55.79)		108 (32.73)
ISUP 1 (3 + 3)	5664 (6.79)	1530 (22.9)	1 (0.37)	349 (21.4)		65 (19.7)
ISUP 2 (3 +4)	2053 (2.46)	538 (8.05)	1 (0.37)	142 (8.71)		43 (13.09)
ISUP 3 (4 + 3)	903 (1.08)	261 (3.91)	2 (0.74)	66 (4.05)		49 (14.85)
ISUP 4 (4 +4, 3 +5 and 5 + 3)	689 (0.83)	424 (6.35)	45 (16.61)	92 (5.64)		19 (5.76)
ISUP 5 (4 +5, 5 +4 and 5 +5)	568 (0.68)	205 (3.07)	221 (81.55)	72 (4.42)		26 (7.88)

Table 1: Subject characteristics among all biopsied men in the STHLM3 study and among men whose biopsies were digitized, tabulated by men **(top)** and by individual biopsy cores **(bottom).** No cancer grade information is shown for Imagebase, as the grading of this set of samples was performed independently by multiple observers. Imagebase cancer length was assessed by L.E.



Avoided benign biopsy cores n (%*)	Detected cancer biopsy cores n (%*)	Missed cores with cancer by ISUP score, n (%*) ISUP 1	ISUP 2	ISUP 3	ISUP 4	ISUP 5	Missed men with cancer n (%*)
▸ 570 (62.6)	720 (99.9)	0 (0.0)	1 (0.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
▸ 788 (86.6)	718 (99.6)	2 (0.6)	1 (0.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
▸ 809 (88.9)	716 (99.3)	4 (1.1)	1 (0.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
▸ 864 (94.9)	714 (99.0)	4 (1.1)	2 (1.4)	0 (0.0)	0 (0.0)	1 (1.4)	1 (0.5)

Avoided benign biopsy cores n (%*)	Detected cancer biopsy cores n (%*)	Missed cores with cancer by ISUP score, n (%*) ISUP 1	ISUP 2	ISUP 3	ISUP 4	ISUP 5	Missed men with cancer n (%*)
▸ 49 (45.4)	221 (99.5)	1 (1.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.8)
▸ 78 (72.2)	220 (99.1)	2 (3.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.8)
▸ 94 (87.0)	219 (98.6)	3 (4.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.8)
▸ 97 (89.8)	217 (97.7)	3 (4.6)	1 (1.6)	1 (2.0)	0 (0.0)	0 (0.0)	1 (1.8)

Figure 1: ROC curves and AUC for cancer detection by individual cores and by men **(left)** with four operating points on the core level curve highlighted **(right)** for the independent test set **(top)** and the external validation set **(bottom)**. The first two columns from left show the number of biopsy cores that could be discarded from further consideration and the number of biopsy cores that would need pathological evaluation, respectively. The values in parentheses indicate the corresponding specificity and sensitivity. The next five columns show the number and percentage of missed malignant cores by ISUP score for each operating point. The rightmost column indicates the number and percentage of missed cancers among all men with cancer.

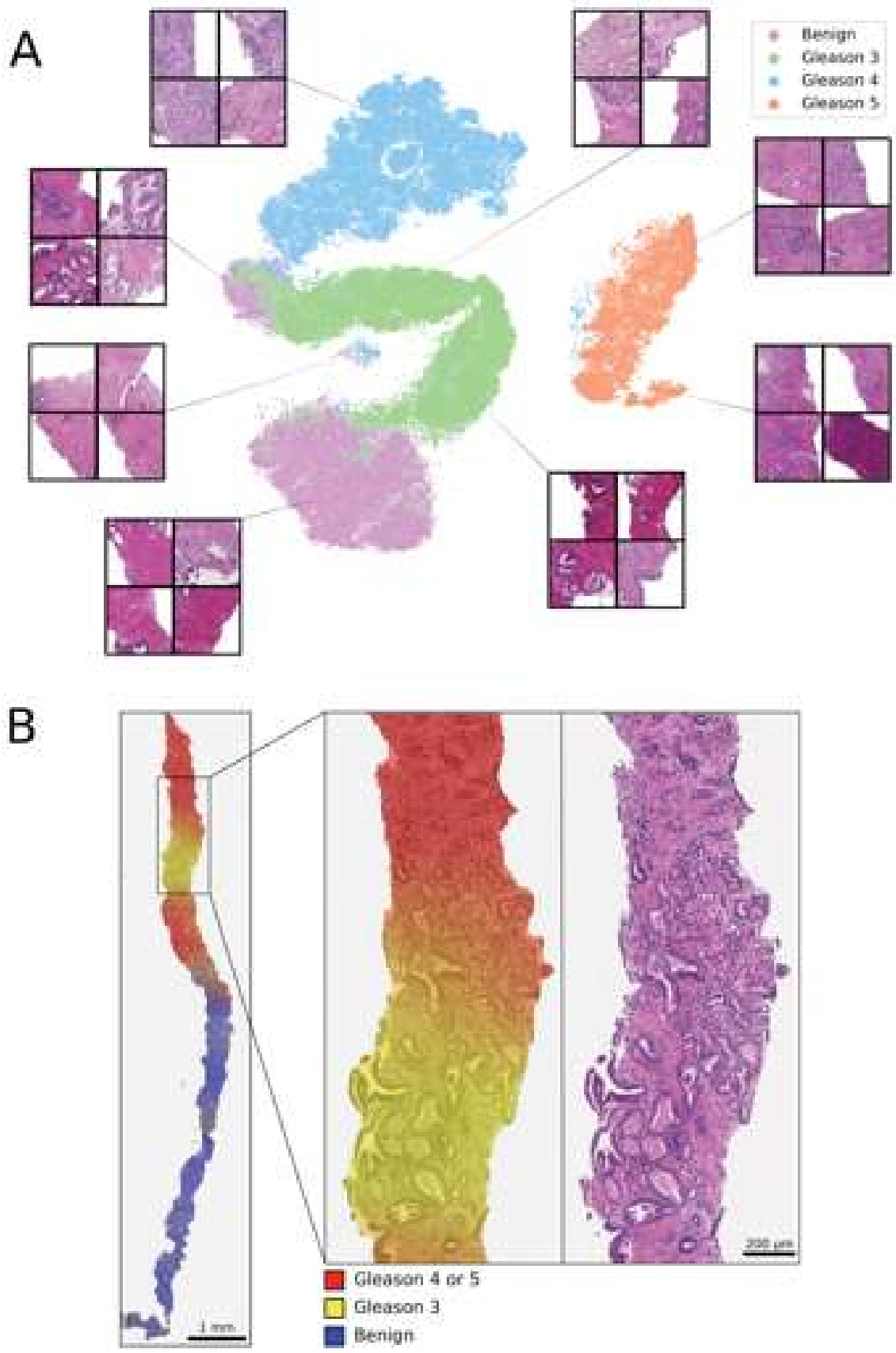


Figure 2: (A) A t-SNE visualization of the feature space learned by the DNN model. Feature representations of patches were obtained by extracting the activations of the penultimate layer of the grading DNN. The 2048-dimensional feature vectors were then reduced to two dimensions using t-SNE. Each data point (n=153,484) represents a single patch from a random sample stratified by grade

(including benign, Gleason 3, Gleason 4, Gleason 5). The colors indicate the grade assigned to each patch by the study pathologist. Four example patches (538 µm x 538 µm), randomly selected from each of the observed clusters, are shown.
(B) Color-coded visualization of cancer grades estimated by the AI. The colors represent the estimated probabilities for the presence of benign (blue), malignant low grade (Gleason 3, yellow) and malignant high grade (Gleason 4 or 5, red) tissue at different locations of the biopsy (left). A magnified view of the AI output (center) and the corresponding H&E stained tissue (right) are shown for a region where an estimated transition between low- and high-grade morphology can be observed. This core from the test data was graded as ISUP 3 (GS 4+3) by the study pathologist.

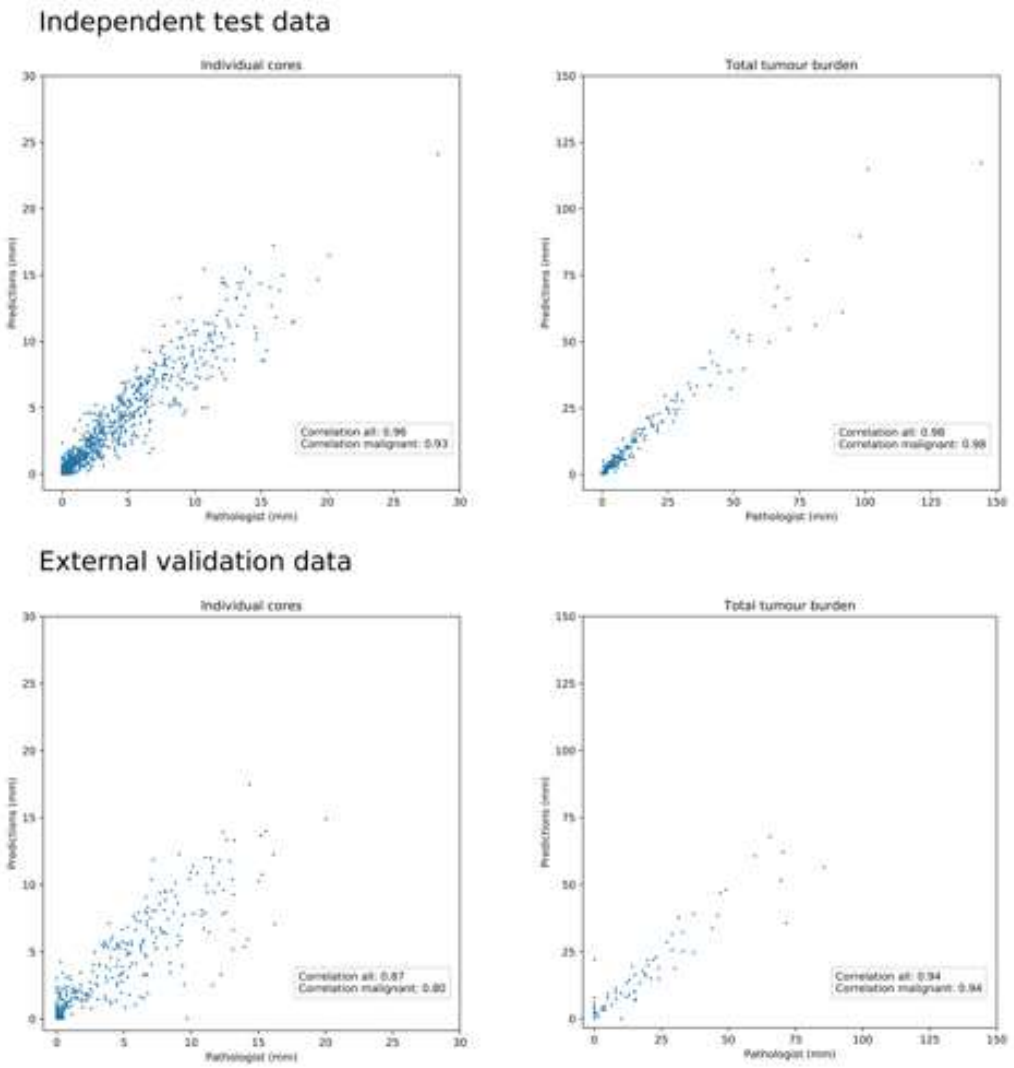


Figure 3: Scatterplots presenting the concordance between cancer lengths estimated by the AI and the pathologist for indepedent test data. Results are shown for individual cores **(left)** and aggregated over cores for each man **(right)** for the independent test set **(top)** and external validation set **(bottom)**. Corresponding linear correlation coefficients computed for all cores and malignant cores only are shown in each plot. Data points in the left plot are jittered along the x-axis for clarity.

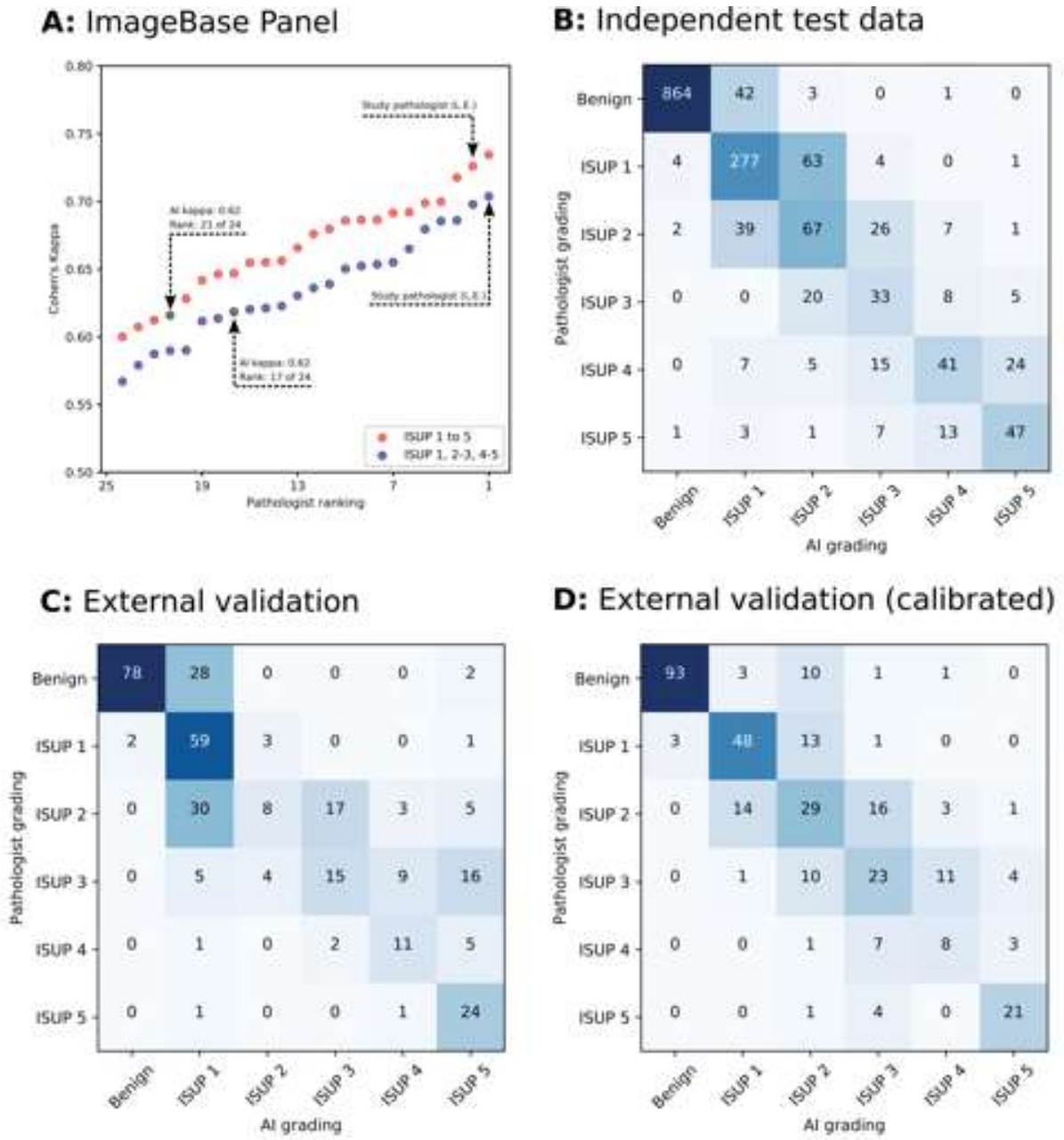


Figure 4: Grading performance on test data. **(A)** Cohen's kappa for each pathologist ranked from lowest to the highest. Each kappa value is the average pair-wise kappa for each of the pathologists compared against the others. To account for the natural order of the ISUP scores we used linear weights. The AI is highlighted with a black dot and an arrow. The study pathologist (L.E.) is highlighted with an arrow. Values computed based on all five ISUP scores are plotted in red, while values based on a grouping of ISUP scores commonly used for treatment decision are shown in blue. **(B)** A confusion matrix on the independent test data of 1631 slides and **(C)** the external validation data of 330 slides. **(D)** Results on external validation data are additionally shown following calibration of the slide-level model. This procedure did not involve any model retraining. The pathologist's (L.E.) grading is shown on the y-axis and the AI's grading on the x-axis. For the independent test set, Cohen's kappa with linear weights was 0.83 when considering all cases, and 0.70 when only considering the cases indicated as positive by the pathologist. For the external validation set, the corresponding values were 0.70 and 0.61. Following calibration, the kappa values increased to 0.76 and 0.66. The results are presented for an operating point achieving a minimum cancer detection sensitivity of 99%.